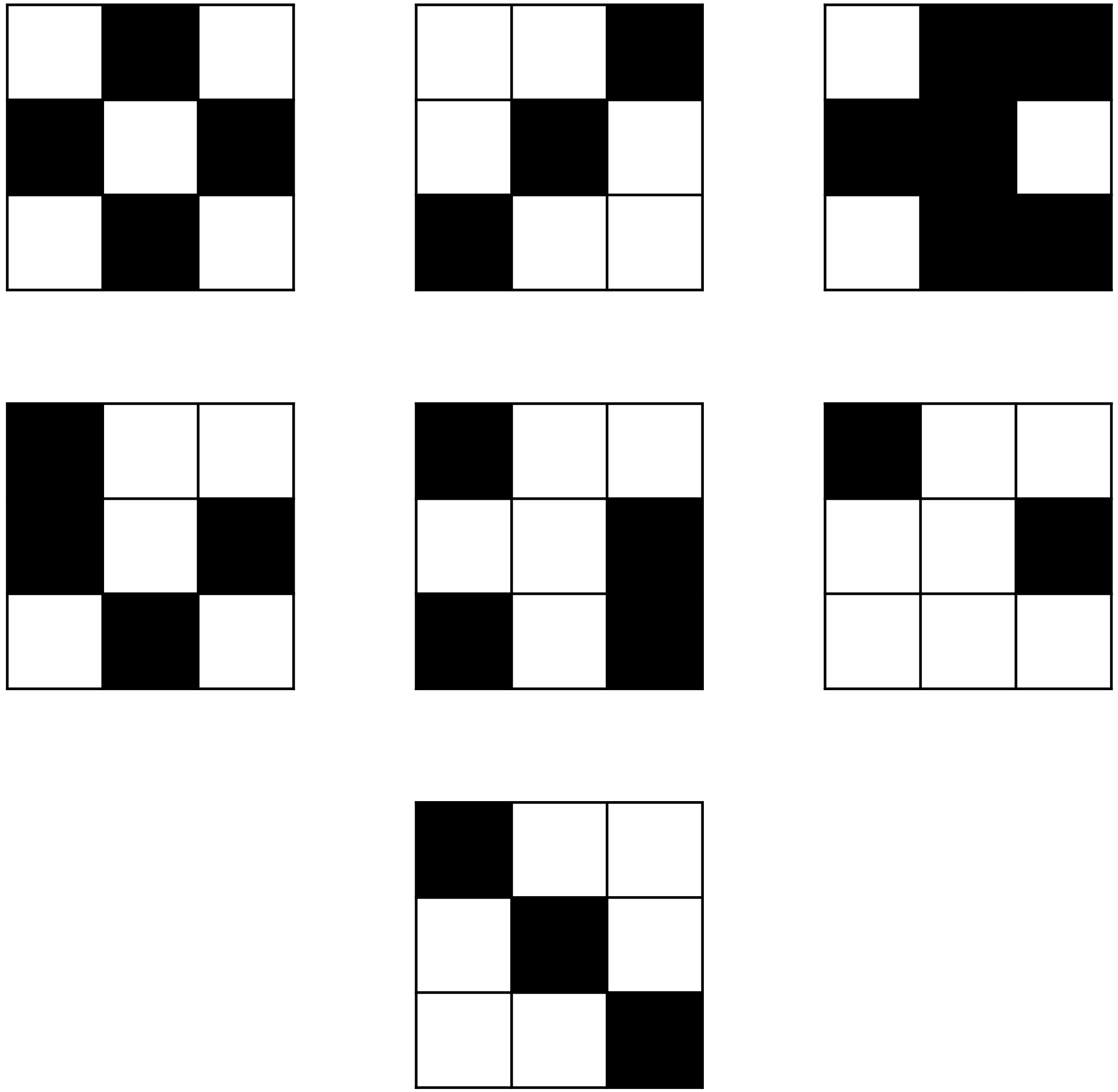
CSE 417T: Introduction to Machine Learning

Lecture 2: Generalization

Henry Chai

08/30/18

Puzzle

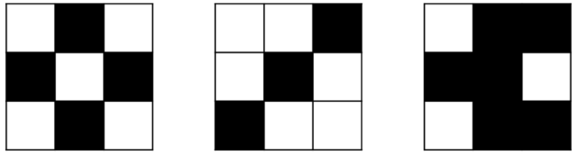


$f(x) = +1$

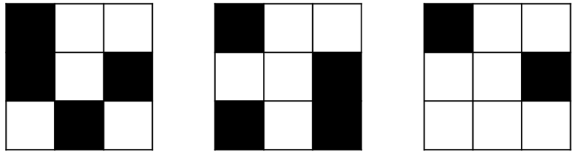
$f(x) = -1$

$f(x) = ???$

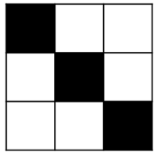
An Answer



$f(x) = +1$



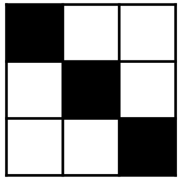
$f(x) = -1$



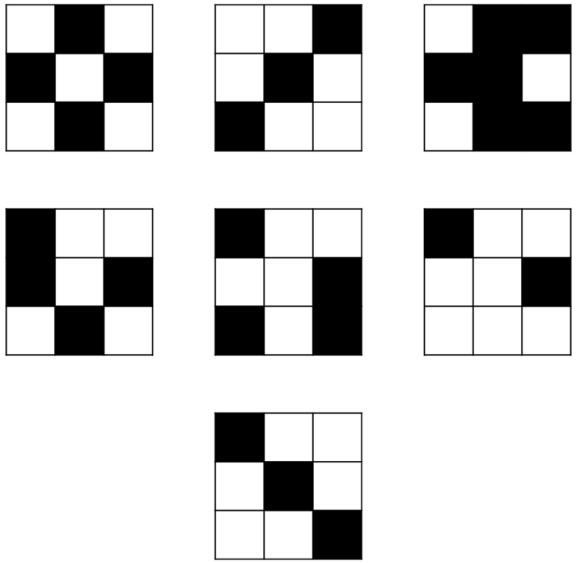
$f(x) = ???$

$$h(x) = \begin{cases} +1 \text{ if symmetric} \\ -1 \text{ otherwise} \end{cases}$$

$$\downarrow$$

$h\Bigg($  $\Bigg) = +1$

An Answer



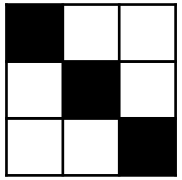
$f(x) = +1$

$f(x) = -1$

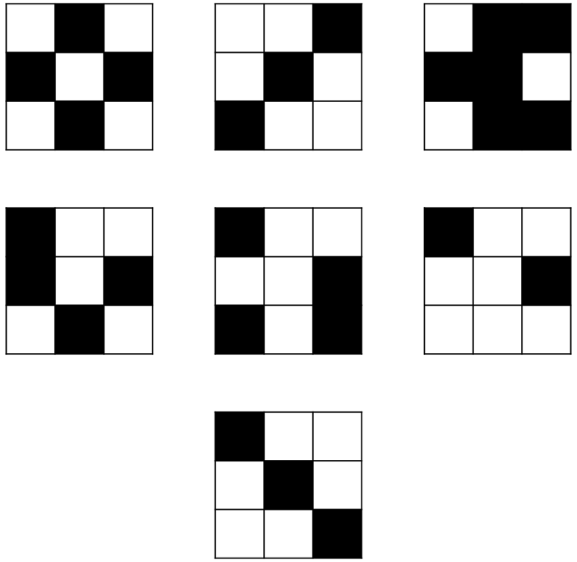
$f(x) = ???$

$$h(x) = \begin{cases} +1 \text{ if top left is white} \\ -1 \text{ otherwise} \end{cases}$$

$$\downarrow$$

$h\Big($  $\Big) = -1$

An Answer

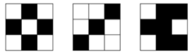


$f(x) = +1$

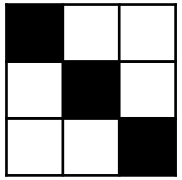
$f(x) = -1$

$f(x) = ???$

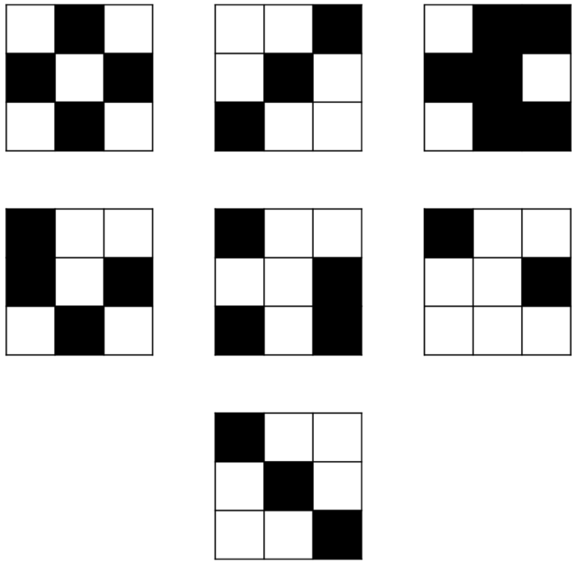
$h(x) = \begin{cases} +1 \text{ if } x \in \{\ \cdot\ \} \\ -1 \text{ otherwise} \end{cases}$



$\downarrow$

$h\big($  $\big) = -1$

An Answer



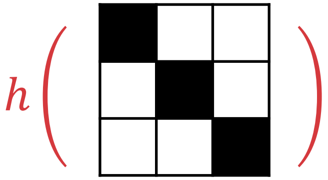
$f(x) = +1$

$f(x) = -1$

$f(x) = ???$

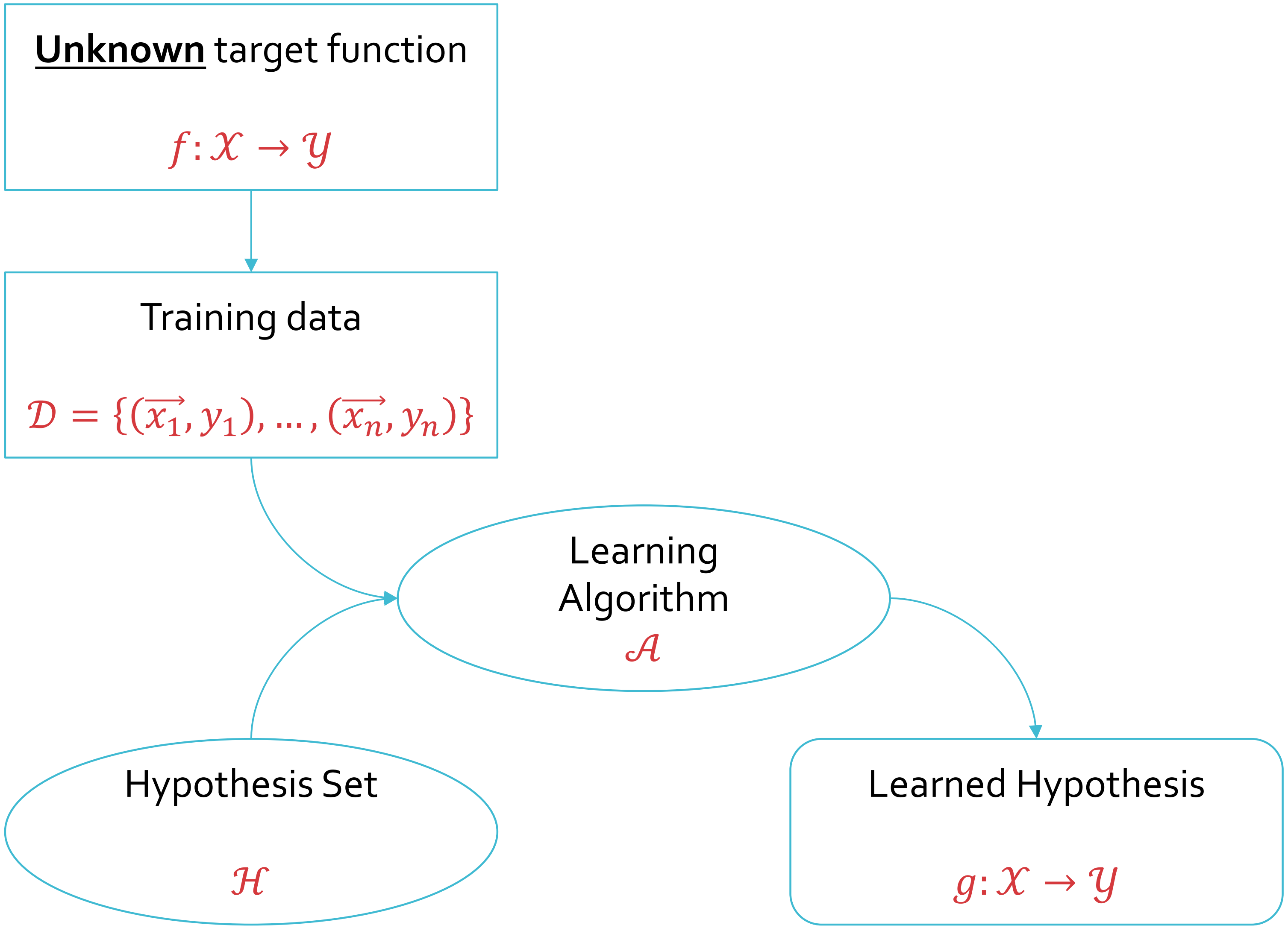
$$h(x) = \begin{cases} +1 \text{ if } x \in \{ \quad \} \\ -1 \text{ otherwise} \end{cases}$$

$\downarrow$

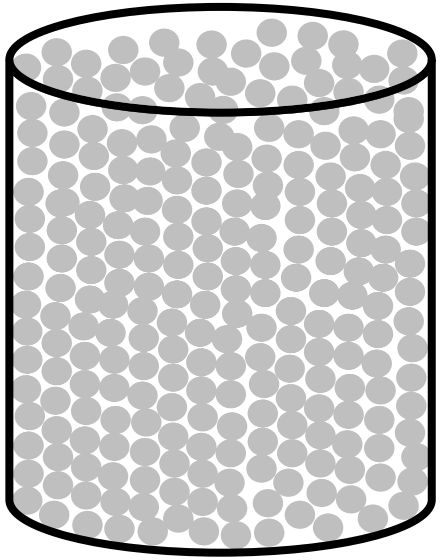


$= +1$

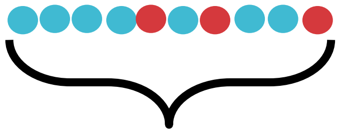
Recall


Unknown target function
$f: \mathcal{X} \rightarrow \mathcal{Y}$
Training data
$\mathcal{D} = \{(\overrightarrow{x_1}, y_1), \ldots, (\overrightarrow{x_n}, y_n)\}$
Learning Algorithm
$\mathcal{A}$
Hypothesis Set
$\mathcal{H}$
Learned Hypothesis
$g: \mathcal{X} \rightarrow \mathcal{Y}$

Analogy



μ = fraction of red marbles in bin



n i. i. d. samples

ν = fraction of red marbles in sample

Does $\nu = \mu$?

Does ν say anything about μ?

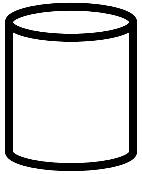
Hoeffding's Inequality

- μ = fraction of red marbles in bin
- ν = fraction of red marbles in a sample of size n

- $P\{|\nu - \mu| > \epsilon\} \leq 2e^{-2\epsilon^2 n}$

- As n increases, RHS decreases
- As ϵ decreases, RHS increases

Connection to Learning

 = input space ($\mathcal{X}$)

 = point in the input space ($\vec{x}$)

 = training data ($\mathcal{D}$)

 = a point classified correctly by a specified hypothesis h

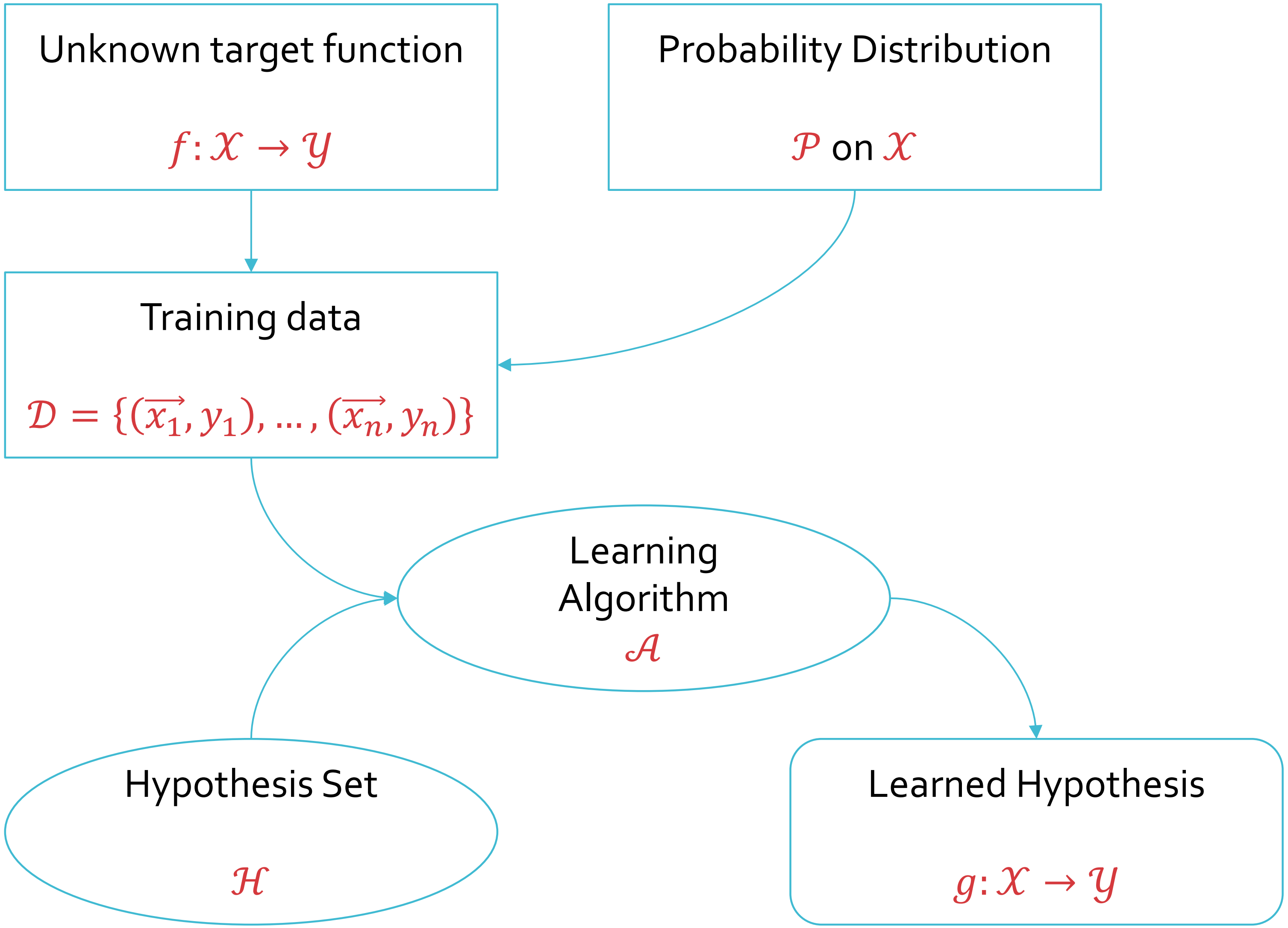
 = a point classified incorrectly by a specified hypothesis h

ν = fraction of training data classified incorrectly by h ($E_{in}(h)$)

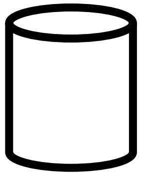
μ = fraction of points in the input space classified incorrectly by h ($E_{out}(h)$)

$$P\{|E_{in}(h) - E_{out}(h)| > \epsilon\} \leq 2e^{-2\epsilon^2 n}$$

Formal Setup



Validation

 = input space ($\mathcal{X}$)

= point in the input space ($\vec{x}$)

 = training data ($\mathcal{D}$)

= a point classified correctly **by a specified hypothesis *h***

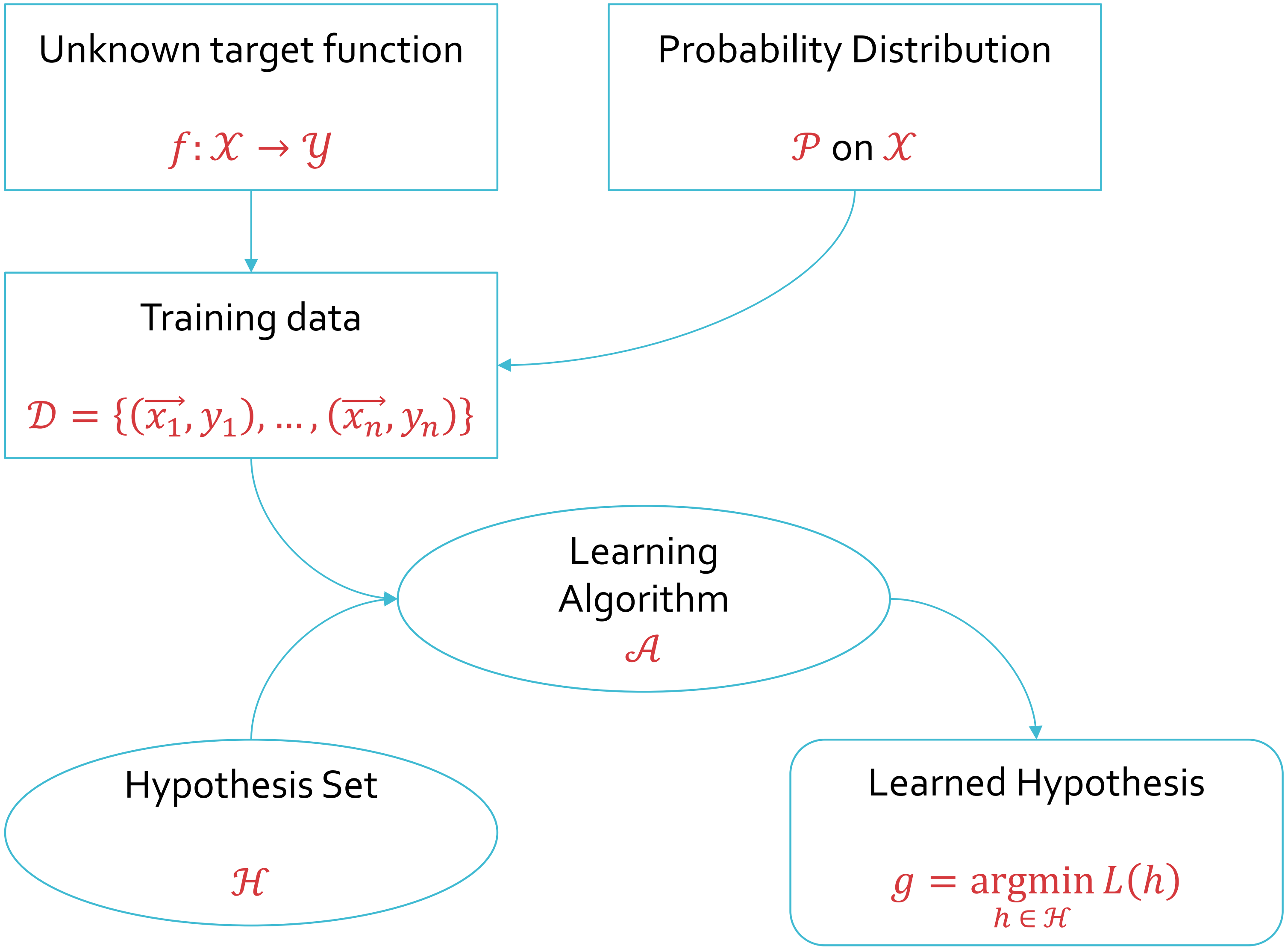
= a point classified incorrectly **by a specified hypothesis *h***

ν = fraction of training data classified incorrectly **by *h*** ($E_{in}(h)$)

μ = fraction of all possible data classified incorrectly **by *h*** ($E_{out}(h)$)

$$P\{|E_{in}(h) - E_{out}(h)| > \epsilon\} \leq 2e^{-2\epsilon^2 n}$$

Formal Setup



Another Analogy

- If you toss a fair coin 20 times, the probability that it comes up heads 20 times is $2^{-20} \approx$ 1e-6
- If you toss 2^{20} fair coins 20 times each, the probability that at least one coin comes up heads 20 times is $\approx 1 - \frac{1}{e} \approx 0.63$

Another Analogy

- Take any coin that came up all heads and apply Hoeffding's inequality

- $P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 2e^{-2\epsilon^2 n}$
- $P\{|\text{Fraction of tails in 20 trials} -$
  $\text{Fraction of tails in } \infty \text{ trials}| > \epsilon\} \leq 2e^{-40\epsilon^2}$
- $P\{|0 - P\{\text{this coin coming up tails}\}| > \epsilon\} \leq 2e^{-40\epsilon^2}$
- $P\left\{P\{\text{this coin coming up tails}\} > \frac{1}{4}\right\} \leq 2e^{-2.5} \approx 0.15$

Hoeffding's Inequality (Corrected)

- Suppose $\mathcal{H}$ is finite i.e. $\mathcal{H} = \{h_1, \dots, h_m\}$

$$P\{|E_{in}(g) - E_{out}(g)| > \epsilon\}$$

$$\leq P\left\{\bigcup_{j=1}^{m} |E_{in}(h_j) - E_{out}(h_j)| > \epsilon\right\}$$

$$\leq \sum_{j=1}^{m} P\{|E_{in}(h_j) - E_{out}(h_j)| > \epsilon\}$$

$$\leq \sum_{j=1}^{m} 2e^{-2\epsilon^2 n} = 2(m)e^{-2\epsilon^2 n}$$

Hoeffding's Inequality (Corrected)

- Suppose $\mathcal{H}$ is finite i.e. $\mathcal{H} = \{h_1, \ldots, h_m\}$
- $E_{in}(g)$ = in-sample error of best hypothesis in $\mathcal{H}$
- $E_{out}(g)$ = out-of-sample error of best hypothesis in $\mathcal{H}$

- $P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 2(m)e^{-2\epsilon^2 n}$

- As n increases, RHS decreases
- As ϵ decreases, RHS increases
- As m increases, RHS increases